$X_1, \dots, X_n \sim \mathcal{D}_1$ w/ cdf F iid.

$Y_1, \dots, Y_m \sim \mathcal{D}_2$, w/ cdf G iid.

Goal: Test $H_0: F = G$ $H_1: F \neq G$

① Estimators

$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{X_i \leq t\}$

$G_m(t) = \frac{1}{m} \sum_{j=1}^{m} \mathbb{1}\{Y_j \leq t\}$

② Pivot / test statistic

$T_{n,m}$

$\psi = \mathbb{1}\{T_{n,m} > s\}$

③ Adjust s

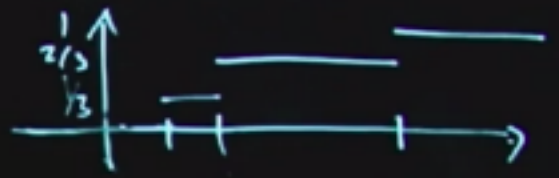
Two-sample Kolmogorov-Smirnov test

Reminder: Kolmogorov-Smirnov goodness of fit test

$X_1, \dots, X_n \sim \mathcal{D}$ w/ cdf F iid.

Test $H_0: F = F_0$ $H_1: F \neq F_0$

$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{X_i \leq t\}$



Empirical cdf

$T_n = \sup_{t \in \mathbb{R}} \sqrt{n} \left| F_n(t) - F_0(t) \right| \xrightarrow[n \to \infty]{\mathcal{D}} Z$

supremum of a Brownian bridge

$X_1, \dots, X_n \sim \mathcal{D}_1$ w/ cdf F iid.

$Y_1, \dots, Y_m \sim \mathcal{D}_2$, w/ cdf iid. G

Assume F continuous & strictly increasing.

Goal: Test $H_0: F = G$ $H_1: F \neq G$

① Estimators

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{X_i \leq t\}$$

$$G_m(t) = \frac{1}{m} \sum_{j=1}^{m} \mathbb{1}\{Y_j \leq t\}$$

② Pivot/ test statistic

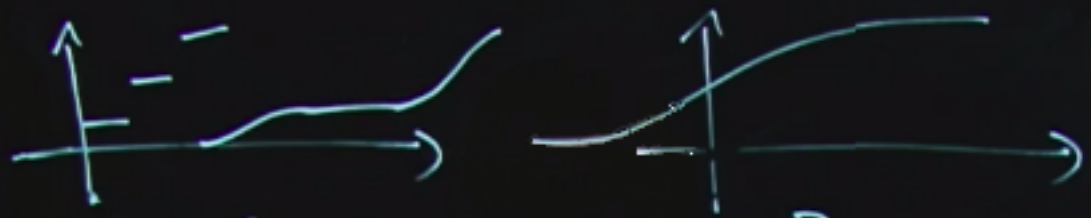
$$T_{n,m} = \sup_{t \in \mathbb{R}} |F_n(t) - G_m(t)|$$

$\psi = \mathbb{1}\{T_{n,m} > s\}$

③ Adjust s

Two-sample Kolmogorov-Smirnov test

② [Glivenko-Cantelli Thm: $\sup_{t \in \mathbb{R}} |F_n(t) - F(t)| \xrightarrow[n \to \infty]{} 0$ a.s.]



$$\{X_i \leq t\} = \{F(X_i) \leq F(t)\}$$

Assp $\Rightarrow \exists F^{-1}: (0,1) \to \mathbb{R}$

$F^{-1}(F(t)) = t \quad \forall t \in \mathbb{R}$

$F(F^{-1}(a)) = a \quad \forall a \in (0,1)$

$X_1, \dots, X_n \sim \mathcal{D}_1$ w/ cdf F iid.

$Y_1, \dots, Y_m \sim \mathcal{D}_2$ w/ cdf G iid.

Assume F continuous & strictly increasing.

Goal: Test $H_0: F = G$ $H_1: F \neq G$

① Estimators

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{X_i \leq t\}$$

$$G_m(t) = \frac{1}{m} \sum_{j=1}^{m} \mathbb{1}\{Y_j \leq t\}$$

② Pivot/ test statistic

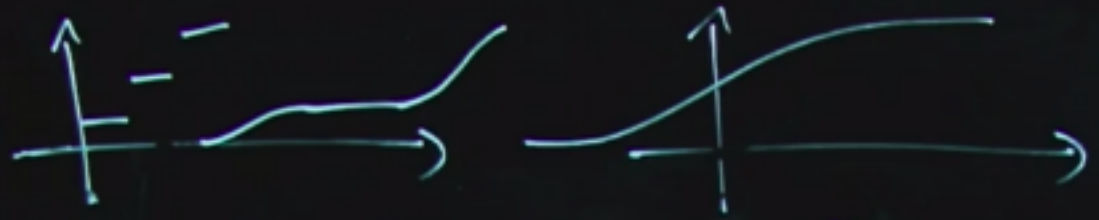
$$T_{n,m} = \sup_{t \in \mathbb{R}} |F_n(t) - G_m(t)|$$

$\psi = \mathbb{1}\{T_{n,m} > s\}$

③ Adjust s

Two-sample Kolmogorov-Smirnov test

② [Glivenko-Cantelli-Thm: $\sup_{t \in \mathbb{R}} |F_n(t) - F(t)| \xrightarrow[n \to \infty]{} 0$ a.s.]



Assp $\Rightarrow \exists F^{-1}: (0,1) \to \mathbb{R}$

$F^{-1}(F(t)) = t \quad \forall t \in \mathbb{R}$

$F(F^{-1}(a)) = a \quad \forall a \in (0,1)$

$\{X_i \leq t\} = \{F(X_i) \leq F(t)\}$

$\mathbb{P}(\{F(X_i) \leq a\}) = \mathbb{P}(X_i \leq F^{-1}(a))$

$= F(F^{-1}(a)) = a \quad \forall a \in (0,1) \Rightarrow F(X_i) \sim \mathcal{U}([0,1])$

$$T_{n,m} = \sup_t \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{X_i \leq t\} - \frac{1}{m} \sum_{j=1}^{m} \mathbb{1}\{Y_j \leq t\} \right|$$

$$= \sup_t \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{\underbrace{F(X_i)}_{=U_i \sim \mathcal{U}([0,1])} \leq F(t)\} - \frac{1}{m} \sum_{j=1}^{m} \mathbb{1}\{\underbrace{F(Y_j)}_{=V_j \sim \mathcal{U}([0,1])} \leq F(t)\} \right|$$

$$= \sup_{a \in (0,1)} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{U_i \leq a\} - \frac{1}{m} \sum_{j=1}^{m} \mathbb{1}\{V_j \leq a\} \right|$$

$X_1, \dots, X_n \sim \mathcal{D}_1$ w/ cdf F iid.

$Y_1, \dots, Y_m \sim \mathcal{D}_2$, w/ cdf G iid.

Assume F continuous & strictly increasing.

Goal: Test $H_0: F = G$ $H_1: F \neq G$

① Estimators

$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{X_i \le t\}$

$G_m(t) = \frac{1}{m} \sum_{j=1}^{m} \mathbb{1}\{Y_j \le t\}$

② Pivot/ test statistic

$T_{n,m} = \sup_{t \in \mathbb{R}} |F_n(t) - G_m(t)| \Rightarrow \mathbb{P}(T_{n,m} > s) \approx \alpha$

$\psi = \mathbb{1}\{T_{n,m} > s\}$

③ Adjust s

Two-sample Kolmogorov-Smirnov test

$$T_{n,m} = \sup_{a \in (0,1)} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{U_i \le a\} - \frac{1}{m} \sum_{j=1}^{m} \mathbb{1}\{V_j \le a\} \right|, \quad U_i, V_j \overset{\text{iid.}}{\sim} \mathrm{Unif}(0,1)$$

③ Non-asymptotic: Sample $\ell = 1, \dots, M$ U_i^ℓ, V_j^ℓ,

calculate $T_{n,m}^\ell = T_{n,m}(U_i^\ell, V_j^\ell)$

$$H_M(q) = \frac{1}{M} \sum_{\ell=1}^{M} \mathbb{1}\{T_{n,m}^\ell \le q\}$$

Pick $s = \hat{q}_{1-\alpha}$, $1-\alpha$ quantile of H_M.

$\approx q_{1-\alpha}$, $1-\alpha$ — u — of cdf of $T_{n,m}$

$X_1, \dots, X_n \sim D_1$ w/ cdf F iid.

$Y_1, \dots, Y_m \sim D_2$, w/ cdf G iid.

Assume F continuous & strictly increasing.

Goal: Test $H_0: F = G$ $H_1: F \neq G$

① Estimators

$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{X_i \le t\}$

$G_m(t) = \frac{1}{m} \sum_{j=1}^{m} \mathbb{1}\{Y_j \le t\}$

② Pivot / test statistic

$T_{n,m} = \sup_{t \in \mathbb{R}} |F_n(t) - G_m(t)|$

$\psi = \mathbb{1}\{T_{n,m} > \delta\}$

③ Adjust δ

Two-sample Kolmogorov-Smirnov test

③ Asymptotic TKS test: $T_{n,m} = \sup_{t \in \mathbb{R}} \left[\sqrt{n} |F_n(t) - F_0(t)| \right]$

Fix t, $\mathrm{Var}(F_n(t)) = \mathrm{Var}\left(\frac{1}{n} \sum_{i=1}^{n} \underbrace{\mathbb{1}\{X_i \le t\}}_{\sim Be(F(t))} \right)$

$= \frac{1}{n^2} \cdot n \cdot F(t)(1 - F(t)) = \frac{1}{n} F(t)(1 - F(t))$

Now, for $T_{n,m}$: $\mathrm{Var}(F_n(t) - G_m(t)) = \mathrm{Var}\left(\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{X_i \le t\} - \frac{1}{m} \sum_{j=1}^{m} \mathbb{1}\{Y_j \le t\} \right)$

$= \frac{1}{n^2} \cdot n F(t)(1 - F(t)) + \frac{1}{m^2} \cdot m F(t)(1 - F(t))$

$= \left(\frac{1}{n} + \frac{1}{m} \right) F(t) \cdot (1 - F(t))$

$\frac{n+m}{n \cdot m} \Rightarrow \tilde{T}_{n,m} = \sqrt{\frac{n \cdot m}{n+m}} \cdot \sup_{t \in \mathbb{R}} |F_n(t) - G_m(t)| \xrightarrow[n,m \to \infty]{D} Z$, sup of Brownian bridge